Application of GAM to Crime Analysis Data

Prof Stan Openshaw and Dr Ian Turton
Centre for Computational Geography
School of Geography
University of Leeds
Leeds LS2 9JT, UK

email: stan,ian @geog.leeds.ac.uk

1 Geographical Analysis Machine

The Geographical Analysis Machine (Openshaw et al 1987, 1988) is an early attempt at automated exploratory spatial data analysis that was easy to understand. The GAM sought to answer a simple practical question; namely given some point referenced data of something interesting WHERE might there be evidence of localised clustering if you do not know in advance where to look due to lack of knowledge of possible causal mechanism or if prior knowledge of the data precluded testing hypotheses on the database. More simply put, here is a geographically referenced database, now tell me if there are any clusters or crime hot spots and if so where are they located. It offers a solution to those researchers and users of GIS who want to perform a fast exploratory geographical analysis of their data with a minimum of effort. It is an automated procedure that is designed to yield safe results that are largely self-evident.

GAM reflects the view that useful spatial analysis tools have to be able to cope with both the special nature of spatial data and end-users who do not have degrees in statistics. The results have also to be easily understood and self-evident so that they can be readily communicated to other non-experts. This need has been clearly expressed as follows: “We want a push button tool of academic respectability where all the heavy stuff happens behind the scenes but the results cannot be misinterpreted” (Adrian Mckeon, Infoshare: email: 1997). There is also a requirement for results expressed as pretty pictures rather than statistics.

The original code required a Cray YMP supercomputer but it will now run on a UNIX workstation or PC. Compute times are a function of the size of data set; for example, 150,000 points took 700 seconds on a Sun workstation. The latest version of the method dates from 1991 but it was revived in 1997 after the results of a comparison with seven other cluster detectors in a rare disease context. A recent report by the International Agency for Research in Cancer in Lyons (France) concluded that GAM/K was shown to be the best or equivalent best means of <u>both</u> TESTING FOR THE PRESENCE OF CLUSTERING and for FINDING THE LOCATIONS OF CLUSTERS. Alexander and Boyle (1996) authors of the IARC study concluded: “The GAM has potential applications in this area if adequate computer resources are available. At the present time, however, the new, more sophisticated version of the GAM is complex, difficult to understand..” (p 157). That was in 1991 and these particular criticisms no longer apply. The key point here is that rare disease data are very hard to analyse. Most of the more general spatial analysis needs in a crime mapping context are far easier. So a method that works well on rare disease data might be expected to perform even better on crime data.

2 How does GAM work?

The GAM algorithm involves the following steps:

Step 1. Read in X,Y data for population at risk and a variable of interest from a GIS
Step 2 Identify the rectangle containing the data, identify starting circle radius, and degree of overlap
Step 3 Generate a grid covering this rectangle so that circles of current radius overlap by the desired amount
Step 4 For each grid-intersection generate a circle of radius r
Step 5 Retrieve two counts for the population at risk and the variable of interest
Step 6 Apply some "significance" test procedure
Step 7 Keep the result if significant
Step 8 Repeat Steps 5 to 7 until all circles have been processed
Step 9 Increase circle radius and return to Step 3 else go to Step 10
Step 10 Create smoothed density surface of excess incidence for the significant circles using a kernel smoothing procedure and aggregating the results for all circles
Step 11 Map this surface

Note that the original GAM/1 consisted of Steps 1 to 9. Steps 10-11 are the GAM/K version

The choice of significance test is not considered as being too critical. The aim is not to test conventional hypotheses but merely determine whether or not an observed positive excess incidence is sufficiently large to be unusual and hence of interest It is more a measure of unusualness or surprise than a formal statistical significance test. A number of different measures of unusualness can be applied depending on the rarity of incidence of interest; e.g. Poisson, binomial, bootstrapped zscores, and Monte Carlo tests based on rates. The aim here is not a formal test of significance, instead 'significance' is being used only as a descriptive filter employed to reject circles. It is the map created by the overall distribution of significant circles that is of most interest.

Finally, some of the critics of GAM argue that any clusters found on the output map could well be the consequence of testing multiple hypotheses. The argument is as follows: If you set some arbitrary significance threshold; e.g. alpha=0.05 and if you test 100 hypotheses then 5 will be false positives (i.e. they will appear as being significant wrongly). If you test 1,000,000 hypotheses then on average 50,000 will appear as false positives. There are two problems with this argument: it assumes the hypotheses are independent whereas in a GAM search they are clearly not (because the circles overlap) and it ignores the geography of the problem as it is surely quite different if all the significant circles occur around one or two locations rather than be scattered randomly all over the map. These effects can be studied by Monte Carlo simulation and this feature is now built into GAM. Quite simply, you re-run the entire process on 500 or 1000 randomly generated crime distributions and compare the results with the observed ones.

3 What did we do to the data?

The original data were converted to 1 m references. GAM has a mechanism for handling data uncertainly; viz. the search circles overlap so there is a constant sensitivity analysis being performed as an integral part of the analysis.

GAM seeks to identify localised hot spots which are defined as an accumulation of excess crimes over and above what might have been expected from the population at risk data. These hot spots are mapped providing a visual feel and clue to the location and strength. The more extreme the values the more unusual the result. Simulation can be used to identify whether the observed hot spots could have occurred by chance. GAM is quite sophisticated in that the population at risk can be modified to handle any covariates believed to be important.

4 Residential crimes

The data was analysed at street block scale as GAM can cope with very large data sets and the supplied data was really small. The best results will come from using the finest spatial resolution of the data. The obvious population at risk is the census population although this could be adjusted to reflect socio-economic covariates. This was not done here.

Figure 1 shows the hot spots detected by GAM in the residential crime data. There are two large hot spots (labelled A and B in the figure). Hot spot B is the larger while hot spot A is strong, its location in a "peninsular" means that edge effects may be responsible for some of the excess, circles located near the edge of the zones may take in crimes located in the edge zones, but attempt to draw population from the missing areas of the map leading to false excesses. It may be interesting in the future to repeat the analysis with central Baltimore included in the study area since this boundary is clearly not a low crime area as with the outer boundaries of the study area.

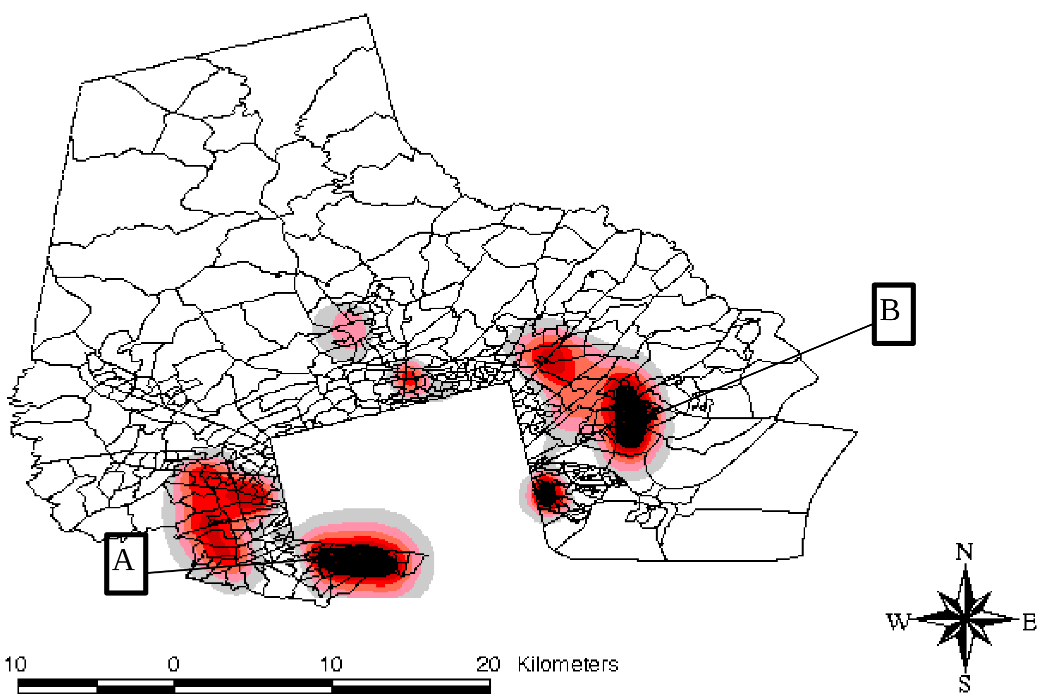

Figure 1: Residential crime hot spots, population at risk census night population

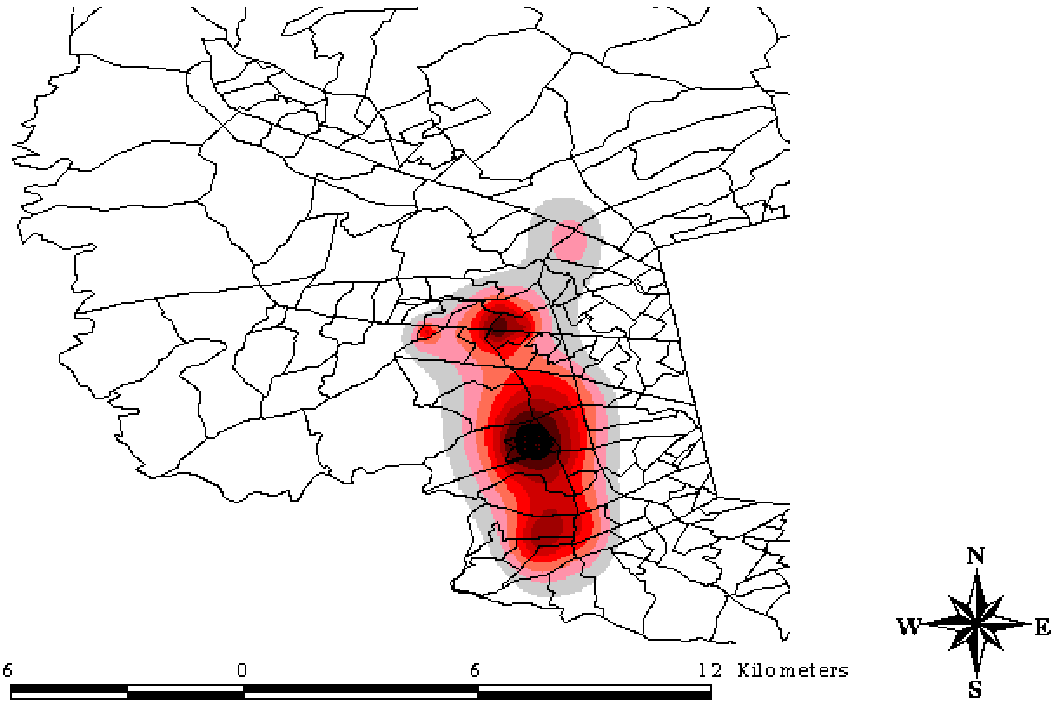

Figure 2: Residential crime hot spots in the small study region.

Figure 2 shows the single hot spot detected in the smaller study area, there are two small hot spots to the north and south of the main central hot spot.

5 Street crimes

There are two possible populations at risk: census night time population and the length of street in each polygon. We investigated both, since it is relatively easy to calculate the length of street in each block group using an intersect command in the GIS.

Figure 3 shows the hot spots detected if resident population is used as the population at risk. Again there are two large hot spots detected however in this case they are both situated well back from the edges of the map.

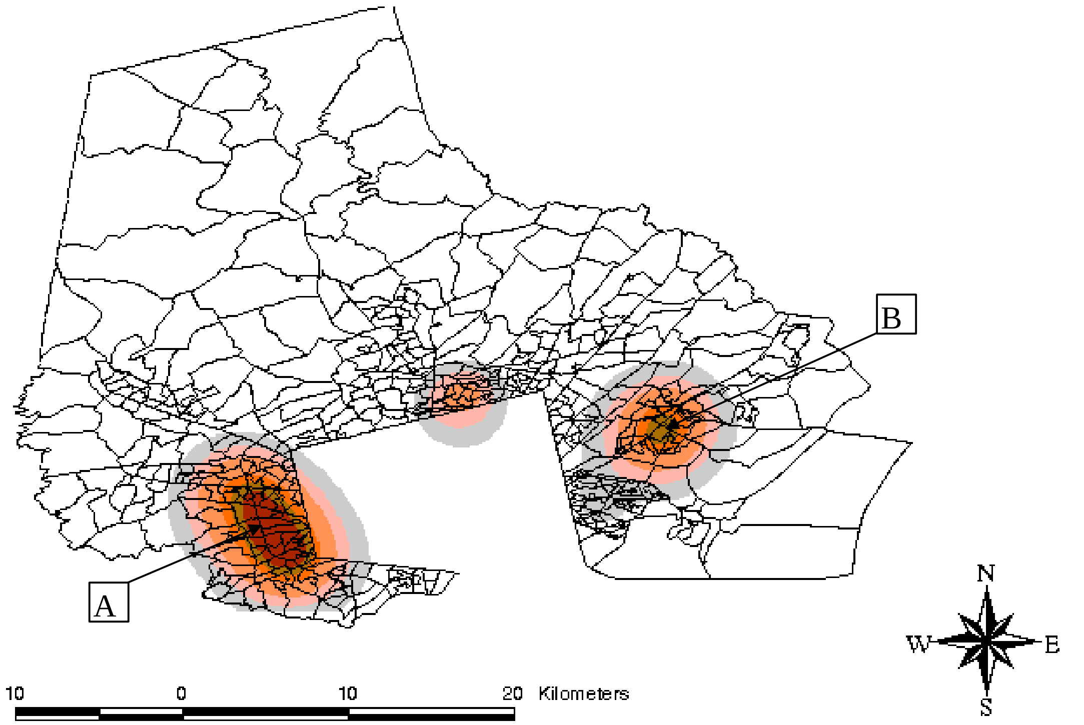

Figure 3: Street Crime Hot spots, population at risk census night population

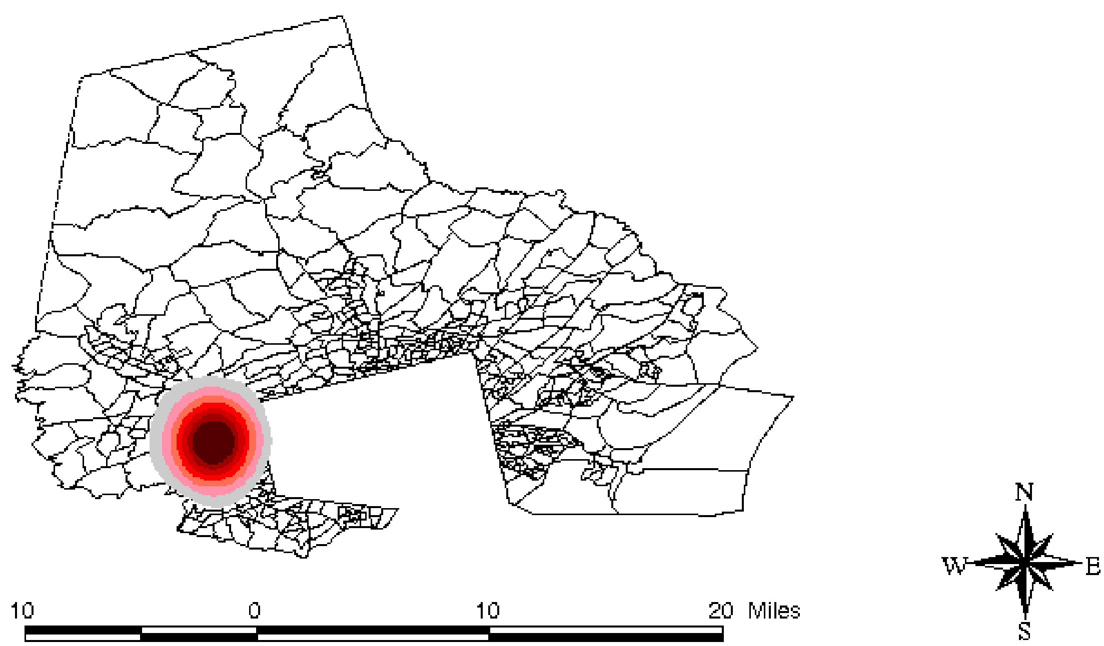

Figure 4: Street crime hot spots, population at risk street length

Figure 4 shows the pattern of hot spots detected by GAM when the length of street in each block group is used to define the population at risk, here only a single hot spot in the south west of the map is detected.

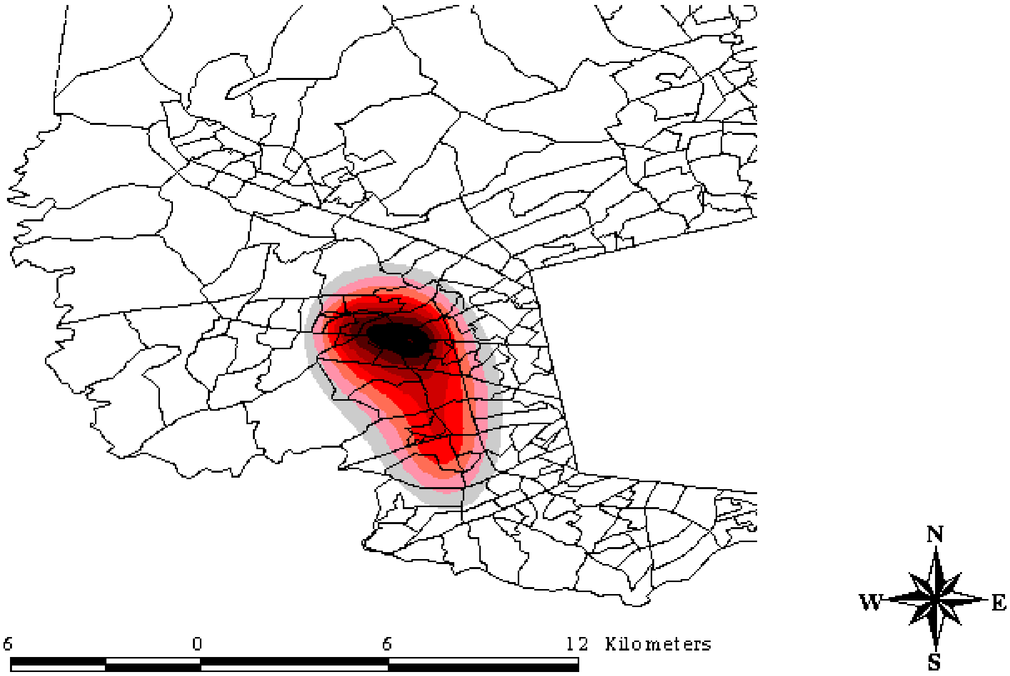

Figure 5: Street crime hot spot in small study region

Figure 5 shows a single hot spot detected with the street crime data was analysed in the smaller study region.

6 Using GAM to do more sophisticated crime data analysis

The data used here are less than adequate. GAM is a cluster detector. It will happily search for clusters in any kind of crime data. The results would probably be most useful if they related to particular time periods, particular times of day, particular modus operandi, if the population at risk had relevant covariates included in it, if the full data set were being analysed rather than a subset with potentially large edge effects. GAM can be used to monitor crime patterns. If you used last years (or last time period's) crime data to define a population at risk you could then measure changes from it. Also because GAM is automatic then it can be used to spot hot spots as they emerge in real-time; most computer runs take only a few minutes.

A residual question is how do you know when a hot spot that appears on the map is not a real hot spot at all? There are four possible answers. One is to feed randomised crime data into GAM and see what sorts of hot spots are found. Typically, they will be weak pathetic things that are not "highly peaked"; but how high does a peak have to be before you get excited about it? The answer is data dependent and qualitative. It also depends on what you plan to use the results for. The safest and simplest strategy is merely to look for the highest peaks. The second answer is to expend lots more compute time on multiple testing by Monte Carlo simulation. This will indicate how easy it is to obtain results as extreme as those being observed by running GAM on multiple sets of randomly generated crime data sets with similar incidence to the observed data. The third solution is to keep for training purposes the results for differing degrees of clustering and use them to train the user to discriminate between the massively interesting and the rubbish. The strength of GAM is that it is a visual method of analysis. It is meant to suggest and create new insights in an almost artistic and qualitative kind of way. The fourth solution is to use GAM purely as an automated procedure and re-install the simple expert system that GAM/K had in 1990 before it was decided to use the human eye-ball instead. It is important to remember that there are many different causes of hot spots, many of which are related to the quality of the data being analysed. No machine based procedure can at present detect "bad data" for you; you have to do it and 100% automation would so reduce the value of human inputs as to render these spurious causes of clustering totally invisible to the end-user.

7 Getting a copy of GAM

We are working on developing downloadable versions of GAM. An on-line test facility and considerable help documentation can be found at http://www.ccg.leeds.ac.uk/smart/intro.html. The on-line version of GAM allows users to view results in their browser, download the results for later analysis in a GIS or move around a VRML model of the results, an example of this is shown in figure 4. There are plans for a Sun Solaris 2.5, Java, and NT versions for later this year

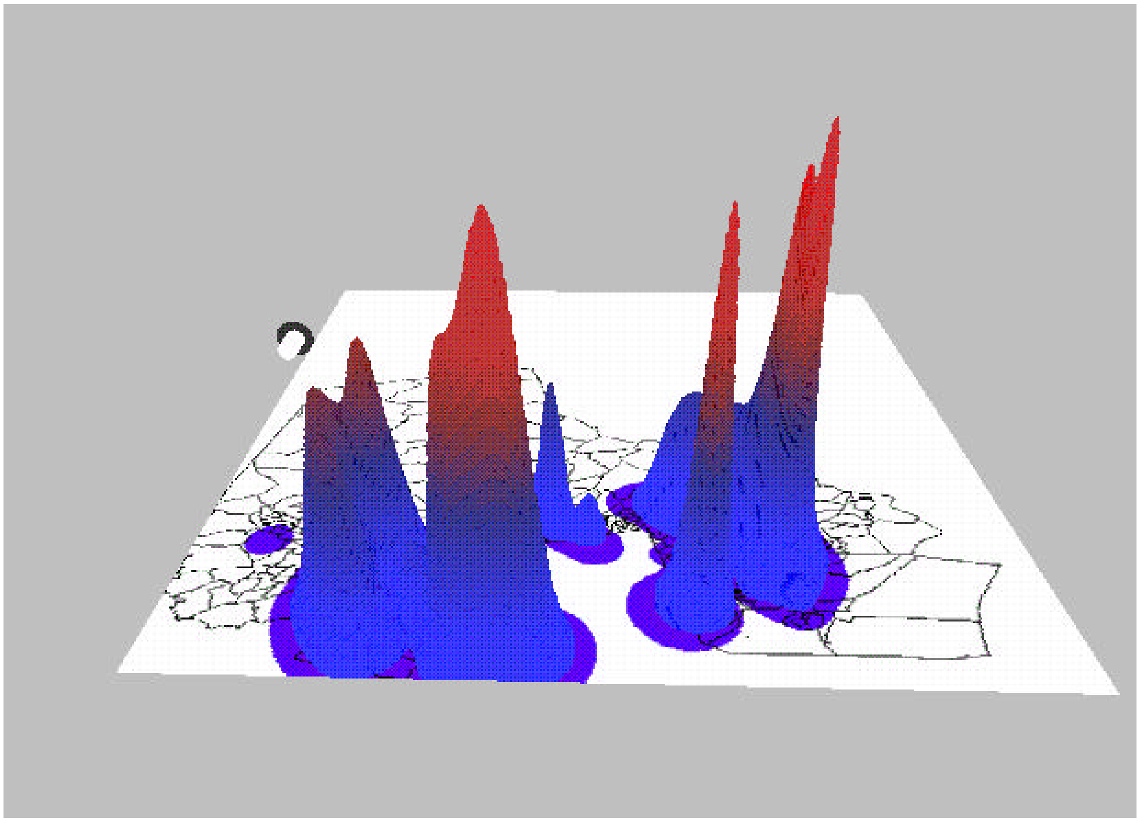

Figure 4: A VRML model of residential crime hot spots.

Acknowledgement.

The research reported was partly supported by the ESRC via Grant R237260.

References

Alexander, F. E., Boyle, P., 1996 Methods for Investigating localised Clustering of Disease IARC Scientific Publications No 135, Lyon, France

Openshaw, S., Charlton, M., Wymer, C. & Craft, A.W., 1987, A mark I geographical analysis machine for the automated analysis of point data sets. Int. J. GIS, 1, 335-358

Openshaw, S., Charlton, M., Craft, A.W. & Birth, J.M., 1988, Investigation of leukaemia clusters by the use of a geographical analysis machine, Lancet, I, 272-273

Openshaw, S., Wilkie, D., Binks, K., Wakeford, R., Gerrard, M.H. & Croasdale, M.R., 1989, A method of detecting spatial clustering of disease. In: Crosbie, W.A. & Gittus, J.H., eds, Medical Response to Effects of Ionising Radiation, London, Elsevier, p.295-308

Openshaw S. & Craft, A. 1991, Using the Geographical Analysis Machine to search for evidence of clusters and clustering in childhood leukaemia and non-Hodgkin lymphomas in Britain. In: Draper, G., ed., The Geographical Epidemiology of Childhood Leukaemia and Non-Hodgkin Lymphoma in Great Britain 1966-83, London, HMSO, p109-122

Openshaw, S., 1996, 'Using a geographical analysis machine to detect the presence of spatial clusters and the location of clusters in synthetic data', in F. E. Alexander and P. Boyle (eds) Methods for Investigating Localised Clustering of Disease IARC Scientific Publication No 135, Lyon, France, p68-87